Inference of functional connectivity in Neurosciences via Hawkes processes

Patricia Reynaud-Bouret
Laboratoire Jean Alexandre Dieudonné
University of Nice Sophia-Antipolis
Nice, France
reynaudb@unice.fr

Vincent Rivoirard
CEREMADE
University Paris-Dauphine
Paris, France
Vincent.Rivoirard@dauphine.fr

Christine Tuleau-Malot
Laboratoire Jean Alexandre Dieudonné
University of Nice Sophia-Antipolis
Nice, France
malot@unice.fr

***Abstract*—We use Hawkes processes as models for spike trains analysis. A new Lasso method designed for general multivariate counting processes [1] enables us to estimate the functional connectivity graph between the different recorded neurons.**

I. Introduction

In Neurosciences, the action potentials (spikes) are the main components for the real-time information processing in the brain. Moreover it is possible to record in vivo several neurons and have access to simultaneous spike trains. Those data are fundamentally random and can be modeled easily by time point processes, i.e. random countable sets of points on $\mathbb{R}_+$. One of the fundamental questions is to guess whether the neurons behave independently or not (see [2]). Such *local dependence* (see [3] for further developments) can be modeled by *multivariate Hawkes processes* [4].

To describe Hawkes models, we need the notion of *conditional intensity* (see e.g. [4]). If a point process N has a conditional intensity $\lambda(.)$, then $t \to \lambda(t)$ is a random predictable function that may depend on the past occurrences of N. Informally, given the past, the quantity $\lambda(t)dt$ gives the conditional probability to have a new occurrence around time t. This means that if we denote by $\mathcal{F}_{t-}$, the past information before time t, we have:

$$\mathbb{E}(dN(t)|\mathcal{F}_{t-}) = \lambda(t)dt, \tag{1}$$

where $dN(t)$ represents the point measure, that is the sum of the Dirac masses at each point of the process N. This notion of "instantaneous expectation given the past" is very useful, because it helps to handle the process in a more intrinsic way, with respect to the classical expectation.

The multivariate Hawkes process models the instantaneous firing rates of M different neurons, with spike trains $N^{(1)}, ..., N^{(M)}$ so that the conditional intensity of the mth point process $N^{(m)}$ is defined for all $t > 0$ by:

$$\lambda^{(m)}(t) = \left(\nu^{(m)} + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(m)}(t-u) dN^{(\ell)}(u) \right)_+, \tag{2}$$

where the $\nu^{(m)}$'s are positive parameters representing the *spontaneous firing rates* and where the $h_\ell^{(m)}$'s are called the *interaction functions* and have support in $\mathbb{R}_+^*$. More precisely, before the first occurrence of the multivariate process, the $N^{(m)}$'s behave like homogeneous Poisson processes with constant intensities $\nu^{(m)}$. The first occurrence (and the next ones) affects all the processes by increasing or decreasing the conditional intensity via the interaction functions $h_\ell^{(m)}$'s.

For instance, if $h_\ell^{(m)}$ takes large positive values in the neighborhood of the delay d and is null elsewhere, then after the delay d of one occurrence of $N^{(\ell)}$, the probability to have a new occurrence of $N^{(m)}$ will significantly increase: the process $N^{(\ell)}$ excites the process $N^{(m)}$. On the contrary, if $h_\ell^{(m)}$ is negative around d, then after the delay d of one occurrence of $N^{(\ell)}$, the probability to have a new occurrence of $N^{(m)}$ will significantly decrease: the process $N^{(\ell)}$ inhibits the process $N^{(m)}$. Note in particular that the functions $h_m^{(m)}$'s model self-interactions.

If one draws an arrow from ℓ to m whenever the interaction function $h_\ell^{(m)}$ is non zero, one can draw a graphical model of local independence as in [3], which, to some extent, should reflect the functional connectivity between the recorded neurons, i.e. this graph reflects the way recorded neurons are interacting. Figure 1 gives some examples of graphical models with only 3 recorded neurons.

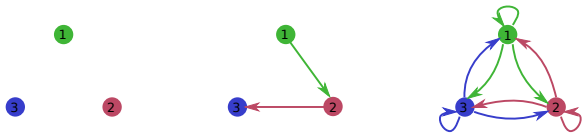

Fig. 1. Examples of graphical models of local independence.

Note that the Hawkes process as described above cannot really model non-stationary data. Indeed, when t grows (and under conditions on the interaction functions), the process converges quite quickly towards an equilibrium, which is stationary (see for instance [5] and the references therein). If those conditions are not satisfied, the number of points in the process grows too fast to be a realistic model for spike trains anyway. Hence Hawkes processes as defined in (2) cannot model inhomogeneous data but can model dependent data.

Therefore, if we observe the spike trains on $[0, T_{max}]$, we fix an interval $[T_1, T_2] \subset [0, T_{max}]$, typically an interval where all the estimated mean firing rates seem constant, where we assume that the processes are obeying a Hawkes model. We

want to estimate on this smaller interval

$$f^* = \left((\nu^{(m)})_{m=1,\dots,M}, (h_\ell^{(m)})_{\ell,m=1,\dots,M} \right), \qquad (3)$$

where it is assumed that the interaction functions are bounded with support contained in $(0, A]$ with $T_1 > A$. Inference for Hawkes models has been studied for a long time, in particular for parametric models, using likelihood [6] and it has been used in Neurosciences in [7]. However, in Neurosciences, it is not possible to design a parametric model with few parameters on the spike trains. Hence in [7] piecewise constants functions with hundreds of parameters were used to estimate the interaction functions. The number of points that need to be observed to perform correctly such estimation is huge (several thousands of points). This amount is usually not reached on the real spike trains data, because the lack of stationnarity cannot be neglected over trials involving that many points. More recently, inference has been obtained by combining likelihood with tests and various kinds of penalization [8], [9] or with Bayesian models on the real-valued parameters and on the graph [10]. Those works used for Hawkes processes, or variations of them, are essentially parametric with no theoretical guarantee when the parametric models are false. More classical model selection based on AIC has also been used in genomics [11]. However it does not adapt well to irregular nonparametric functions. This is the reason why adaptive inference has recently been developed in such models. The univariate case ($M = 1$) has been studied in [12], where a model selection method based on ℓ_0 penalty is developed. However the high computational cost for the multivariate analysis makes this method quite useless in practice. Hence a multivariate approach has been developed in [1] for more general multivariate processes, based on a weighted ℓ_1 penalty. In the sequel, we detail this last method and apply it to the spike trains analysis.

II. Intensity candidates and least-squares contrast

We first define conditional intensity candidates. Let $\mathcal{H}$ be the Hilbert space of all $f = (\mu^{(m)}, (g_\ell^{(m)})_{\ell=1,\dots,M})_{m=1,\dots,M}$ such that the $\mu^{(m)}$'s are real parameters and the $g_\ell^{(m)}$'s are bounded functions with support in $(0, A]$. For any f in $\mathcal{H}$, we consider the predictable linear transformation $\psi(f) = (\psi^{(1)}(f), ..., \psi^{(M)}(f))$ such that for any $t > 0$,

$$\psi_t^{(m)}(f) = \mu^{(m)} + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} g_\ell^{(m)}(t-u) dN^{(\ell)}(u). \qquad (4)$$

Note that $\lambda^{(m)}(t) = [\psi_t^{(m)}(f^*)]_+$. Therefore $\psi^{(m)}(f)$ can be considered as a good intensity candidate as long as it is close enough to $\lambda^{(m)}$ and this even if it becomes slightly negative. In this context, the following quantity $D(f, f^*)$ measures somehow the distance between f and f^*:

$$D^2(f, f^*) := \sum_{m=1}^{M} \int_{T_1}^{T_2} [\psi_t^{(m)}(f) - \lambda^{(m)}(t)]^2 dt. \qquad (5)$$

Of course, we cannot compute it without knowing f^*. But minimizing it with respect to f is equivalent to minimizing

$$\sum_{m=1}^{M} \int_{T_1}^{T_2} \left(-2\psi_t^{(m)}(f)\lambda^{(m)}(t) + [\psi_t^{(m)}(f)]^2 \right) dt. \qquad (6)$$

By (1), this should be not far from

$$\gamma(f) = \sum_{m=1}^{M} \left[-2 \int_{T_1}^{T_2} \psi_t^{(m)}(f) dN^{(m)}(t) + \int_{T_1}^{T_2} [\psi_t^{(m)}(f)]^2 dt \right],$$

which is the least-squares contrast for point processes with conditional intensity. This observable expression can be minimized if f depends on a finite number of parameters.

If n i.i.d. trials are recorded, each trial i corresponds to the observation of $N_i = (N_i^{(1)}, ..., N_i^{(M)})$, the multivariate Hawkes process whose intensity is given by the predictable transformation denoted ψ_i. Furthermore, to each trial i, we can associate an intensity λ_i and a contrast $\gamma^{(i)}$. The global least-squares contrast over the n trials is then defined by

$$\gamma_n(f) = \sum_{i=1}^{n} \gamma^{(i)}(f). \qquad (7)$$

We use the following notations: for any predictable processes $H = (H_i^{(1)}, ..., H_i^{(M)})_{i=1,\dots n}$, $K = (K_i^{(1)}, ..., K_i^{(M)})_{i=1,\dots,n}$, set

$$H \bullet N = \sum_{i=1}^{n} \sum_{m=1}^{M} \int_{T_1}^{T_2} H_{i,t}^{(m)} dN_i^{(m)}(t), \qquad (8)$$

$$H \diamond K = \sum_{i=1}^{n} \sum_{m=1}^{M} \int_{T_1}^{T_2} H_{i,t}^{(m)} K_{i,t}^{(m)} dt, \qquad (9)$$

and $H^{\diamond 2} = H \diamond H$.

Our method is based on the *dictionary approach*. We choose a dictionary Φ of known functions of $\mathcal{H}$ and we only consider linear combinations of functions of Φ for estimating f^*:

$$f_a = \sum_{\varphi \in \Phi} a_\varphi \varphi, \quad \text{for} \quad a \in \mathbb{R}^\Phi. \qquad (10)$$

Then, by linearity of ψ, one can rewrite (7) as

$$\gamma_n(f_a) = -2a'b_n + a'G_n a, \qquad (11)$$

where for any φ and $\tilde{\varphi}$ in Φ,

$$(b_n)_\varphi = \psi(\varphi) \bullet N \quad \text{and} \quad (G_n)_{\varphi,\tilde{\varphi}} = \psi(\varphi) \diamond \psi(\tilde{\varphi}).$$

Here a' denotes the transpose of a. Note that the vector b_n and the matrix G_n are both data-driven quantities so are observable. Minimizing $a \mapsto \gamma_n(f_a)$ leads to the solution

$$\hat{a}_n = G_n^{-1} b_n, \qquad (12)$$

and the least-squares estimate of f^* is $\hat{f}_n = f_{\hat{a}_n}$.

III. LASSO ESTIMATE

Besides the same drawbacks that least-squares estimates share with MLE, their components are all non-zero almost surely. Therefore the reconstructed functional connectivity graph is complete and not informative with respect to other sparse graphs. Sparsity can be obtained by combining $\hat{f}_n$ with ℓ_1-penalization. Given a vector of positive weights d, the *Lasso estimate* of f^* is $\tilde{f}_n := f_{\tilde{a}_n}$ where $\tilde{a}_n$ is a minimizer of the following ℓ_1-penalized least-square contrast:

$$\tilde{a}_n \in \arg\min_{a\in\mathbb{R}^\Phi}\{-2a'b_n + a'G_n a + 2d'|a|\}. \qquad (13)$$

We can prove the following oracle inequality on $\tilde{f}_n$.

Theorem 1. *We introduce the following two events:*

$$\Omega_{V,B} = \{\forall\,\varphi\in\Phi, \sup_{t\in[T_1,T_2],m,i}|\psi_{i,t}^{(m)}(\varphi)| \le B_\varphi \text{ and } (\psi(\varphi))^2 \bullet N \le V_\varphi\},$$

for positive deterministic constants B_φ and V_φ and

$$\Omega_c = \left\{\forall\, a\in\mathbb{R}^\Phi, \quad a'G_n a \ge c\; a'a\right\},$$

for a positive constant c. Let x and ε be strictly positive constants and for all $\varphi\in\Phi$,

$$d_\varphi = \sqrt{2(1+\varepsilon)\hat{V}_\varphi^\mu x} + \frac{B_\varphi x}{3}, \qquad (14)$$

with

$$\hat{V}_\varphi^\mu = \frac{\mu}{\mu-\phi(\mu)}\;(\psi(\varphi))^2\bullet N + \frac{B_\varphi^2 x}{\mu-\phi(\mu)}$$

for a real number μ such that $\mu > \phi(\mu)$, where $\phi(\mu) = \exp(\mu)-\mu-1$. Then, with probability larger than

$$1 - 4\sum_{\varphi\in\Phi}\left(\frac{\log\left(1+\frac{\mu V_\varphi}{B_\varphi^2 x}\right)}{\log(1+\varepsilon)}+1\right)e^{-x} - \mathbb{P}((\Omega_{V,B}\cup\Omega_c)^c),$$

the following inequality holds

$$[\psi(\tilde{f}_n)-\lambda]^{\diamond 2} \le C\inf_{a\in\mathbb{R}^\Phi}\left\{[\psi(f_a)-\lambda]^{\diamond 2} + \frac{1}{c}\sum_{\varphi\in S(a)} d_\varphi^2\right\},$$

where C is an absolute positive constant and where $S(a)$ is the support of a, i.e. its coordinates with non-zero coefficients.

The proof is an easy adaptation of Theorem 2 of [1], which is originally stated for $n=1$, to the case of n multivariate Hawkes processes (see also [13]).

This oracle inequality is stated by using a distance between predictable processes expressed via $\diamond$. On the event $\left\{\forall\, i,m,t \quad \lambda_i^{(m)}(t) > 0\right\}$, by linearity of ψ, this can also be seen as a random distance between $\tilde{f}_n$ and f^*. The upper bound is quite classical with a bias term and a variance term, where the "variance" of each coefficient a_φ is measured via d_φ^2. The leading term in d_φ is indeed $\sqrt{2\alpha_{\varepsilon,\mu}\tilde{V}(\varphi)x}$ where $\alpha_{\varepsilon,\mu}$ is strictly larger than 1 but as close to 1 as desired and where $\tilde{V}(\varphi) := (\psi(\varphi))^2\bullet N$ is an unbiased estimate of the bracket of the compensated $(b_n)_\varphi$, the analog of the variance in the martingale setting. The shape of the d_φ's comes from a particular Bernstein type inequality for martingales, stated in [1] and the resulting estimate is therefore called a Bernstein-Lasso estimate.

The variance term is renormalized by c and therefore c gives the rate of convergence of the oracle inequality. It has been proved in [1] where the case $n=1$ is investigated, that if the process is stationary and if the interaction functions are non negative, under some technical assumptions on both the process and the dictionary, then c is of the order of T_2-T_1 with high probability. In [13], the n trials set-up is considered and it has been proved that even if the process is not stationary and under technical assumptions, then the order of magnitude of c is n with high probability. In practice, since G_n is a symmetric non negative observable matrix, it is easy to compute its smallest eigenvalue to assess the quality of the estimate.

It is also quite easy to find V_φ and B_φ such that $\mathbb{P}(\Omega_{V,B})$ is large. With respect to T_2-T_1 and n, their orders of magnitude are typically $\log(T_2-T_1)$ or $\log(n)$ (see [1], [13]).

IV. SIMULATIONS

To illustrate our purpose, let us simulate the very simple graph of local independence corresponding to the second graph of Figure 1, with spontaneous rates $\nu^{(m)} = 10$ Hz for $m=1,2,3$ and interaction functions $h_1^{(2)} = h_2^{(3)} = 160\times\mathbf{1}_{[0.005,0.01]}$, the others being null. We simulate 100 trials with $T_{max}=2$ seconds each. The classical statistical spike trains analysis generally consists in performing various independence tests in the spirit of [2] and on this example, spike trains are declared dependent, but we are not able to distinguish between the second and the third graphical models of Figure 1. Another classical analysis (see Figure 2) consists in performing cross-correlograms between each pair of neurons, i.e to plot the histograms of the $(x_i - x_j)$ for $x_i\in N_i$ and $x_j\in N_j$. There is clearly a favored delay of roughly 5-10 ms between N_1 and N_2 and between N_2 and N_3, giving a visual hint for interaction of the type $1\to 2$ and $2\to 3$, even if it seems that interaction $2\to 3$ is much stronger than $1\to 2$. This difference can be explained: there are more points in N_2 than in N_1, which implies mechanically a larger number of pairs in the second cross-correlogram with respect to the first one. Moreover, there is also, mechanically, a favored delay of roughly 10-20 ms between N_1 and N_3, making practitioners believe that interaction $1\to 3$ exists and seems almost as high as $1\to 2$. This phenomenon is an artefact of cross-correlograms. Since cross-correlograms can only study pairs of neurons, they will see the "convolution" of $h_1^{(2)}$ by $h_2^{(3)}$ as an interaction $1\to 3$. The two presented estimators $\hat{f}_n$ and $\tilde{f}_n$ are able to disentangle the different phenomena and, for $\tilde{f}_n$ to reconstruct the correct graph. To do so, let us restrict ourself to $T_1 = 1$s and $T_2 = 2$s. The dictionnary, Φ, used here, is very simple and consists in regular histograms. Namely, we estimate each interaction function independently by piecewise constant functions on a regular partition of $[0, 0.03]$ of 30 bins. Since there are also 3 spontaneous parameters to estimate,

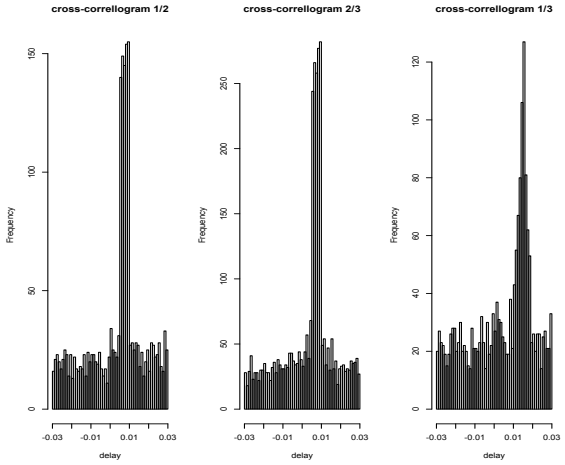

Fig. 2. Cross-correlograms on the simulated data.

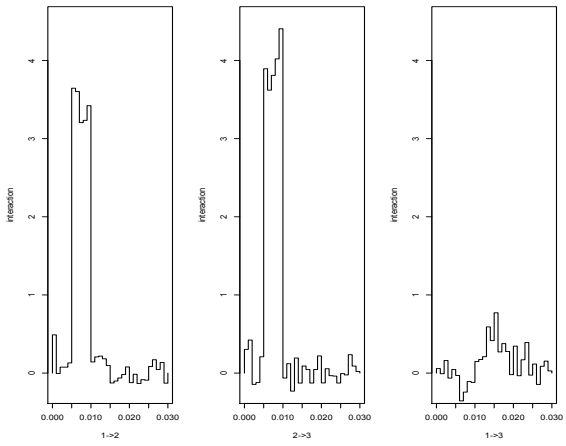

Fig. 3. Reconstructions of $h_1^{(2)}$, $h_2^{(3)}$ and $h_1^{(3)}$ by $\hat{f}_n$.

there are finally 273 parameters to estimate. Figure 3 gives the reconstruction of 3 of the 9 interaction functions by the least-square estimate $\hat{f}_n$. Interactions $1 \to 2$ and $2 \to 3$ have clearly the same strength and the interaction $1 \to 3$ is much more negligible. However with $\hat{f}_n$, the complete graph (i.e. the third graph in Figure 1) is always obtained and there is no clear way to declare that the bump around 10-20 ms for the interaction $1 \to 3$ is significative or not. For a correct sparse reconstruction of the underlying graph of local dependence, we consider $\tilde{f}_n$ (see Figure 4). In this simulation, the weights d_φ are given by

$$d_\varphi = \sqrt{2[(\psi(\varphi))^2 \bullet N]x} + \frac{(\sup_{t\in[a,b],m,i} |\psi_{i,t}^{(m)}(\varphi)|)x}{3}, \quad (15)$$

with $x = \ln(n(T_2 - T_1))$. This formula is simpler than (14) and corresponds to the limit choice $\alpha_{\varepsilon,\mu} = 1$. Figure 4 shows that the support recovery of the interaction functions is perfect. This also implies that the reconstruction of the functional connectivity graph (second graph of Figure 1) is perfect. However as usual for Lasso estimates, shrinkage effects lead to non-negligible bias on the interaction functions that is reduced when performing the ordinary least-square estimate, $\tilde{f}_n^{OLS}$, on the support given by the Lasso estimate $\tilde{f}_n$.

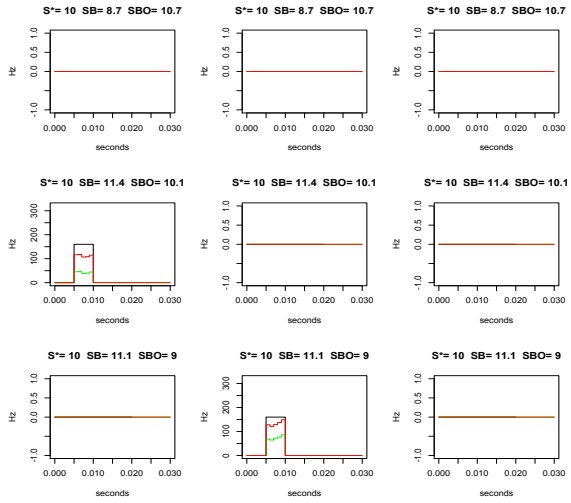

Fig. 4. Reconstructions of $h_\ell^{(m)}$ (ℓ=column, m=row) for $\ell, m = 1, 2, 3$ by $\tilde{f}_n$ in green (**SB** is the spontaneous rate estimate) and by $\tilde{f}_n^{OLS}$ in red (**SBO** is the spontaneous rate estimate). The true f^* appears in black (**S*** is the true spontaneous rate).

A more extensive study can be found in [1] and an application on real neuronal data has been done in [13].

ACKNOWLEDGEMENT

This research is partly supported by the french ANR *Calibration* and the PEPS BMI *Estimation of dependence graphs for thalamo-cortical neurons and multivariate Hawkes processes*.

REFERENCES

[1] N. R. Hansen, P. Reynaud-Bouret, V. Rivoirard, "Lasso and probabilistic inequalities for multivariate point processes.", to appear in Bernoulli.
[2] S. Grün, M. Diesmann, A.M. Aertsen, "Unitary Events Analysis.", *In Analysis of Parallel Spike Trains*, Grün, S., & Rotter, S., Springer Series in Computational Neuroscience, 2010.
[3] V. Didelez, "Graphical models for marked point processes based on local independence."*Journal of the Royal Statistical Society*, Series B, vol. 70, pp. 245-264, 2008.
[4] D.J. Daley, D. Vere-Jones, *An introduction to the theory of point processes., vol. 1* Springer series in statistics, 2005.
[5] P. Brémaud, L. Massoulié, "Stability of nonlinear Hawkes processes.", *Ann. Prob.*, vol. 24(3), pp. 1563-1588, 1996.
[6] T. Ozaki, "Maximum likelihood estimation of Hawkes' self-exciting point processes.", *Ann. Inst. Statist. Math.*, vol. 31(B), pp. 145-155, 1979.
[7] E.S. Chornoboy, L.P. Schramm, A.F. Karr, "Maximum likelihood identification of neural point process systems.", *Biological Cybernetics*, vol. 59, pp. 265-275, 1988.
[8] M. Okatan, M.A. Wilson, E.N. Brown, "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity.", *Neural Computation*, vol. 17, pp. 1927-1961, 2005.
[9] J.W. Pillow, J. Shlens, L. Paninski, A. Sher, A.M. Litke, E.J. Chichilnisky, E.P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population.", *Nature*, vol. 454, pp. 995-999, 2008.
[10] C. Blundell, K.A. Heller, J.M. Beck, "Modelling reciprocating relationships with Hawkes processes.", proceedings of NIPS 2012.
[11] G. Gusto, S. Schbath, "FADO: a statistical method to detect favored or avoided distances between motif occurrences using the Hawkes' model.", *Statistical Applications in Genetics and Molecular Biology*, vol. 4(1), Article 24, 2005.
[12] P. Reynaud-Bouret, S. Schbath, "Adaptive estimation for Hawkes processes; application to genome analysis.", *Ann. Statist.*, vol. 38(5), pp. 2781-2822, 2010.
[13] P. Reynaud-Bouret, V. Rivoirard, F. Grammont, C. Tuleau-Malot, "Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis.", submitted, Hal, 2013.